Introduction to Bayesian Methods - 1

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Course webpage:

http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Bayes.html

Conditional probabilities and Bayes' Theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Joint probability and conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

Conditional probabilities and Bayes' Theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Joint probability and conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem: a purely logical statement

$$P(H|D) = \frac{P(D|H)}{P(D)}P(H)$$

Bayes' theorem again: now as an inferential statement

Likelihood

Prior distribution

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

Posterior distribution

Evidence

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

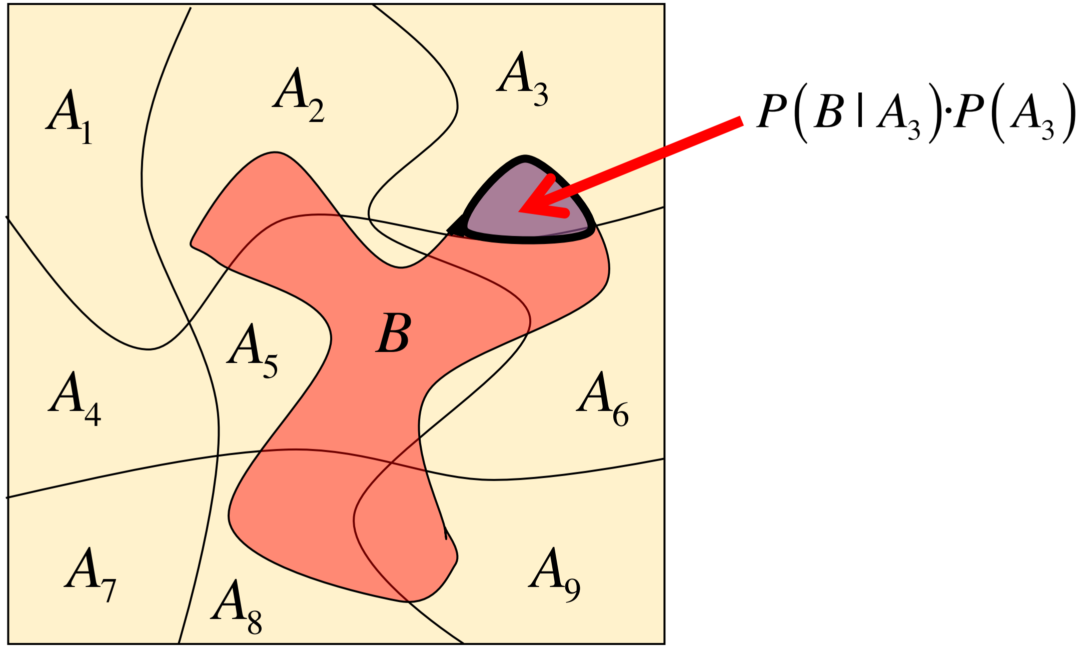
$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{P(B)} \qquad k = 1, \dots, N$$

if the events A_k are mutually exclusive, and they fill the probability space

$$P(B) = P(B \cap \Omega) = P\left(B \cap \bigcup_{n=1}^{N} A_n\right)$$

$$= P\left(\bigcup_{n=1}^{N} (B \cap A_n)\right) = \sum_{n=1}^{N} P(BA_n)$$

$$= \sum_{n=1}^{N} P(B|A_n)\, P(A_n)$$



$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$+$$

$$P(B) = \sum_{k=1}^{N} P(B|A_k) \cdot P(A_k)$$

$$\Downarrow$$

$$P(A_k|B) = \frac{P(B|A_k)}{\sum_{j=1}^{N} P(B|A_j)P(A_j)} P(A_k)$$

Likelihood

Prior distribution

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j)P(H_j)} P(H_k)$$

Posterior distribution

Evidence

MAP estimates

A problem of male twins (Efron, 2003)

Pregnant with twins:
fraternal or identical?

Fraternal: 2/3 of all cases

Identical: 1/3 of all cases

What is the probability of identical twins IF both boys in sonogram?

Answer provided by Bayes theorem

$$P(\text{Identical}|\text{Both boys}) = \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})} P(\text{Identical})$$

$$\left.\begin{aligned} P(\text{Identical}) &= 1/3 \\ P(\text{Fraternal}) &= 2/3 \end{aligned}\right\}$$

Prior probabilities

$$\left.\begin{aligned} P(\text{Both boys}|\text{Identical}) &= 1/2 \\ P(\text{Both boys}|\text{Fraternal}) &= 1/4 \end{aligned}\right\}$$

Conditional probabilities from simple counting argument

$$\begin{aligned} P(\text{Both boys}) &= P(\text{Both boys}|\text{Identical})P(\text{Identical}) \\ &\quad + P(\text{Both boys}|\text{Fraternal})P(\text{Fraternal}) \\ &= (1/2)(1/3) + (1/4)(2/3) = 1/3 \end{aligned}$$

$$\boxed{\begin{aligned} P(\text{Identical}|\text{Both boys}) &= \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})} P(\text{Identical}) \\ &= \frac{(1/2)}{(1/3)}(1/3) = 1/2 \end{aligned}}$$

A simple application to medical tests (HIV testing)

$$P(\text{positive}|\text{infected}) = 1; \qquad P(\text{positive}|\text{not infected}) = 0.015$$

what is the probability *P(infected|positive)* ?

A common answer is 98.5% ... and it is wrong!

Let's use Bayes' theorem ... $P(H_k|D) = \dfrac{P(D|H_k)}{\sum_{n=1}^N P(D|H_n)\, P(H_n)} P(H_k)$

$$\begin{aligned} P(\text{infected}|\text{positive}) &= \frac{P(\text{positive}|\text{infected}) \times P(\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \\ &= \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})}\right] \times P(\text{infected}) \end{aligned}$$

The estimate depends on the size of the infected population

i.e., on the probabilities

P(infected) P(not infected)

$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})}\right] \times P(\text{infected})$$

The posterior estimate strongly depends on prior probability

Example: AIDS testing

(data from https://en.wikipedia.org/wiki/List_of_countries_by_HIV/AIDS_adult_prevalence_rate, accessed May 7th 2022)

$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})}\right] \times P(\text{infected})$$

$$P_{\text{Italy}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.003 + 0.015 \times 0.997} \times 0.003 \approx 16.7\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.173 + 0.015 \times 0.827} \times 0.173 \approx 93.3\%$$

The large number of false positives and the small probability of finding a sick person mean that the probability of being infected if positive is not actually very high.

Repeating measurements changes the reference population.

We incorporate a new positive result in a repeated measurement by using the previous posterior as the new prior:

$$P_{\text{Italy}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.167 + 0.015 \times 0.833} \times 0.167 \approx 93.0\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.933 + 0.015 \times 0.067} \times 0.933 \approx 99.9\%$$

The first test changes the reference population, and the second test, if positive, gives a significant result.

Comparing hypotheses

$$P\left(H_k \mid D,I\right) = \frac{P\left(D \mid H_k,I\right)}{P\left(D \mid I\right)} \cdot P\left(H_k \mid I\right)$$

$$\frac{P\left(H_k \mid D,I\right)}{P\left(H_n \mid D,I\right)} = \overbrace{\underbrace{\left(\frac{P\left(D \mid H_k,I\right)}{P\left(D \mid H_n,I\right)}\right)}_{\text{Bayes' factor}} \cdot \left(\frac{P\left(H_k \mid I\right)}{P\left(H_n \mid I\right)}\right)}^{\text{Odds ratio}}$$

When prior probabilities are the same (equally probable hypotheses), the posterior probability ratio depends only on the Bayes' factor:

$$\frac{P(H_k|D,I)}{P(H_n|D,I)} = \overbrace{\underbrace{\frac{P(D|H_k,I)}{P(D|H_n,I)}}_{\textit{Bayes' factor}} \times \frac{P(H_k|I)}{P(H_n|I)}}^{\textit{Odds ratio}} \underset{\text{Uniform priors}}{=} \underbrace{\frac{P(D|H_k,I)}{P(D|H_n,I)}}_{\textit{Bayes' factor}}$$

From discrete sets of hypothesis to the continuum.
The Bayes' theorem in the context of parameter estimation.

$$P(H_k|D,I) = \frac{P(D|H_k,I)}{P(D|I)} P(H_k,I) = \frac{P(D|H_k,I)}{\sum_{n=1}^{N} P(D|H_n,IU)\; P(H_n,I)} P(H_k,I)$$

$$p\left(\boldsymbol{\theta}|D,I\right) = \frac{P\left(D|\boldsymbol{\theta},I\right)}{\int_{\Theta} P\left(D|\boldsymbol{\theta}',I\right) p\left(\boldsymbol{\theta}'|I\right) d\boldsymbol{\theta}'} \times p\left(\boldsymbol{\theta}|I\right)$$

Likelihood

Prior distribution

Posterior distribution

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

Evidence

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j) P(H_j)} P(H_k)$$

$$p(\theta|D,I) = \frac{P(D|\theta,I)}{\int_\Theta P(D|\theta',I) p(\theta'|I) d\theta} p(\theta|I)$$

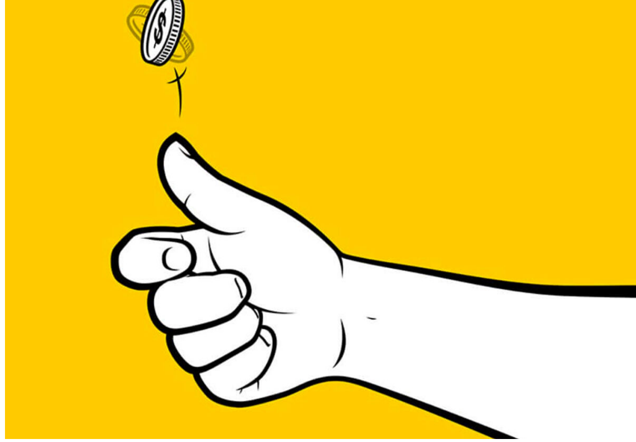
MAP estimates

Consider the following sequence of coin tosses

H, T, T, T, H, H, T, T, H, H, T, H, H, H, H, T, H, H, H, T

(12 heads, 8 tails)

Is this an unbiased coin?

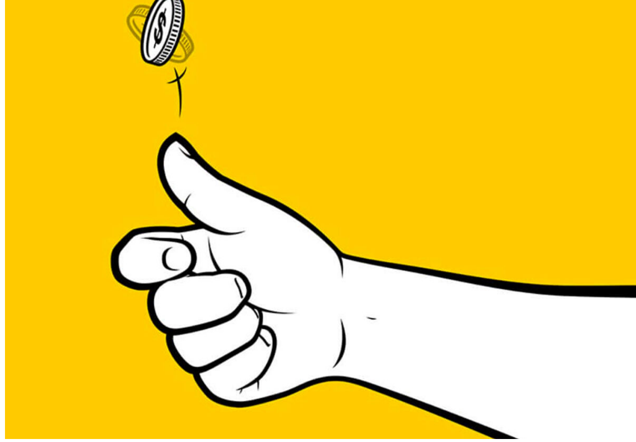


What about the following one?

H, T, H, H, H, T, H, H, T, H, H, H, H, H, T, T, H, T, T, H

(13 heads, 7 tails)

Consider the following sequence of coin tosses



H, T, T, T, H, H, T, T, H, H, T, H, H, H, H, T, H, H, H, T

(12 heads, 8 tails)

Is this an unbiased coin?

Frequentist answer: sample average is 0.6 instead of 0.5, which is just a little less that one standard deviation (≈ 0.11) away from the mean for an unbiased coin

What about the following one?

H, T, H, H, H, T, H, H, T, H, H, H, H, H, T, T, H, T, T, H

(13 heads, 7 tails)

Frequentist answer: sample average is 0.65 instead of 0.5, which is just about 1.36 standard deviations away from the mean for an unbiased coin and is more likely to point to a biased coin

Example of Bayesian inference:

estimate of the (probability) parameter of the binomial distribution

$$P(n|\theta, N) = \binom{N}{n}(1-\theta)^{N-n}\theta^n$$

this is the parameter
that we want to infer
from data

$$p(\theta|n, N) = \frac{P(n|\theta, N)}{\int_0^1 P(n|\theta', N)p(\theta')d\theta'}\, p(\theta)$$

uniform distribution: the
least informative prior

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\ p(\theta')d\theta'}\, p(\theta) = \frac{(1-\theta)^{N-n}\theta^n}{\int_0^1 (1-\theta)'^{N-n}\theta'^n\ d\theta'}$$

the final result is a beta distribution

$$p(\theta|n,N) = \frac{(1-\theta)^{N-n}\theta^n}{\int_0^1 (1-\theta)'^{N-n}\theta'^n \, d\theta'} = \frac{(1-\theta)^{N-n}\theta^n}{B(n+1, N-n+1)}$$

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

beta function

$$\mathrm{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

beta pdf

$$\Rightarrow \quad p(\theta|n,N) = \mathrm{Beta}(\theta|n+1, N-n+1)$$

Mathematical digression: the connection between gamma and beta function

$$\Gamma(a)\Gamma(b) = \int_0^\infty s^{a-1}e^{-s}ds \int_0^\infty t^{b-1}e^{-t}dt$$

$$s = x^2,\ t = y^2 \quad \Rightarrow \quad \Gamma(a)\Gamma(b) = 4\int_0^\infty x^{2a-1}e^{-x^2}dx \int_0^\infty y^{2b-1}e^{-y^2}dy$$

$$x = r\cos\theta,\ y = r\sin\theta \quad \Rightarrow \quad \Gamma(a)\Gamma(b) = 4\int_0^\infty r^{2a+2b-1}e^{-r^2}dr \int_0^{\pi/2} \cos^{2a-1}\theta \sin^{2b-1}\theta d\theta$$

$$(t = \cos^2\theta,\ dt = -2\cos\theta\sin\theta d\theta) \qquad = \Gamma(a+b)\left(2\int_0^{\pi/2} \cos^{2a-1}\theta \sin^{2b-1}\theta d\theta\right)$$

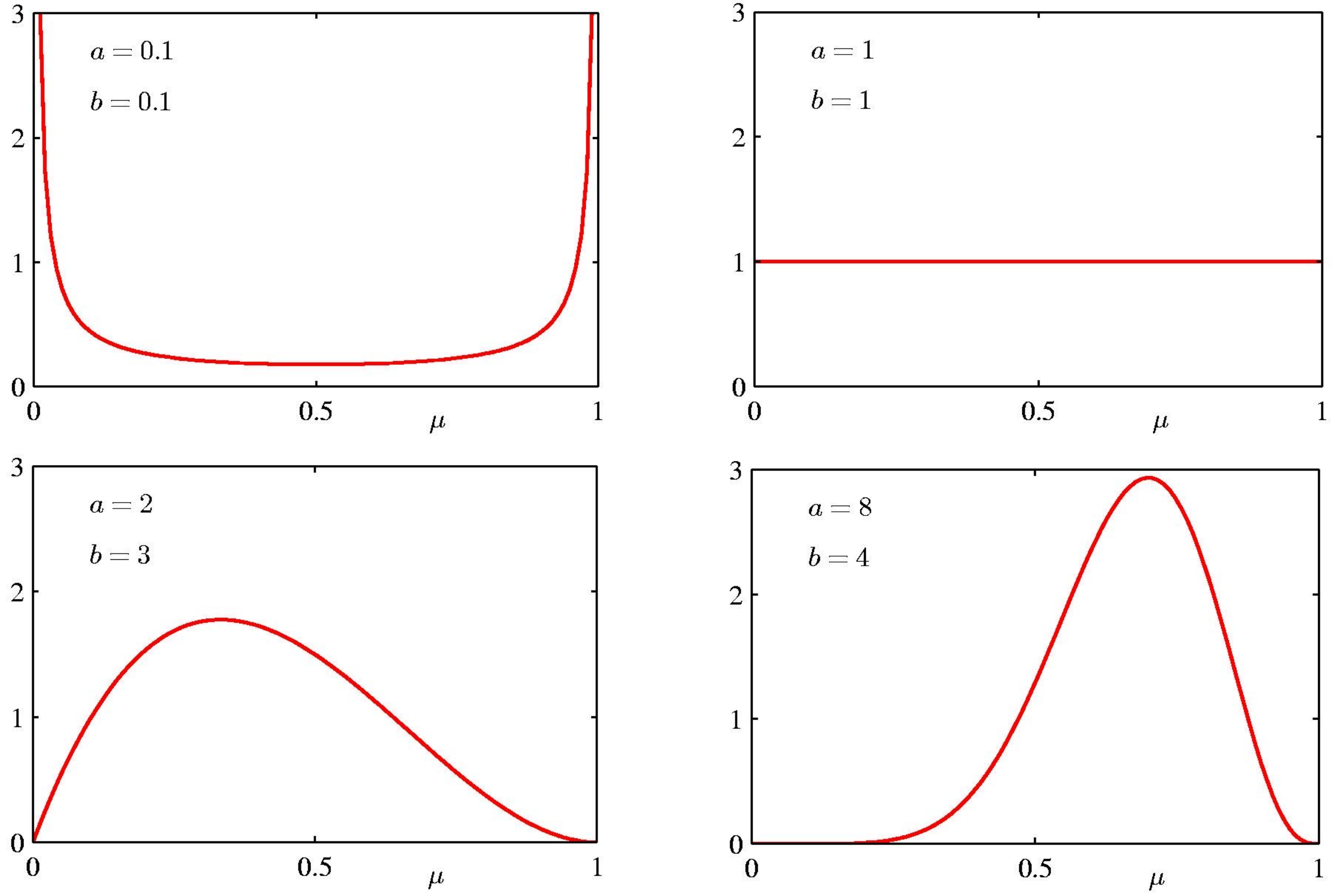
$$= \Gamma(a+b)\int_0^1 t^{a-1}(1-t)^{b-1}dt$$

$$= \Gamma(a+b)B(a,b)$$

$$\Rightarrow \qquad \boxed{B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}$$

Plots of the beta distribution for various values of the hyperparameters a and b.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$



Main statistics of the beta distribution as a function of the a and b **hyperparameters** (a and b, integers)

$$\mathrm{Beta}(\mu|a,b) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a-1}(1-\mu)^{b-1}$$

$$\begin{aligned}
\mathbb{E}(\mu) &= \int_0^1 \mu \; \mathrm{Beta}(\mu|a,b) d\mu \\
&= \int_0^1 \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a}(1-\mu)^{b-1} d\mu \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\mathrm{B}(a+1,b) \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\frac{a!(b-1)!}{(a+b)!} = \frac{a}{a+b}
\end{aligned}$$

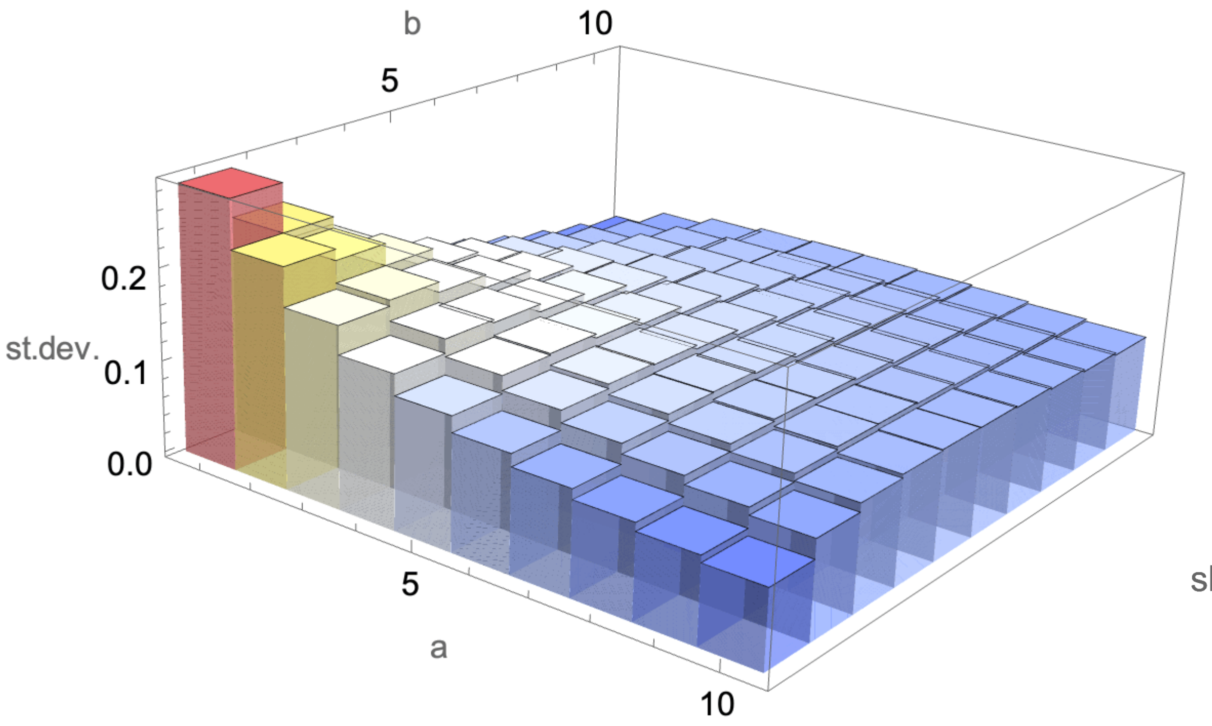
$$\mathrm{var}(\mu) = \mathbb{E}(\mu^2) - \mathbb{E}(\mu)^2 = \frac{ab}{(a+b+1)(a+b)^2}$$

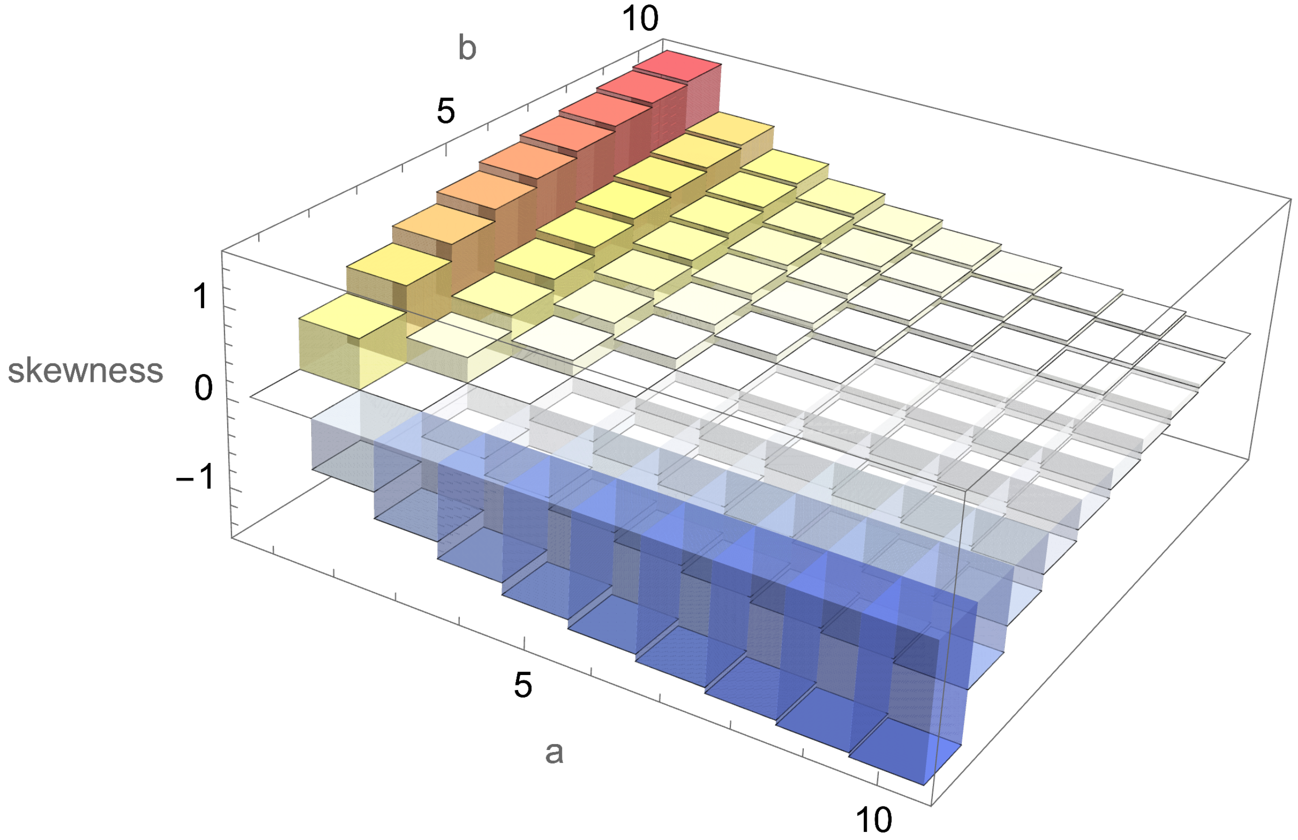
$$\mathrm{skewness}(\mu) = \frac{\mathbb{E}[(\mu-\mathbb{E}(\mu))^3]}{\sigma^3} = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$$

$$\begin{aligned}
\mathbb{E}(\mu^2) &= \int_0^1 \mu^2 \; \mathrm{Beta}(\mu|a,b) d\mu \\
&= \int_0^1 \frac{(a+b-1)!}{(a-1)!(b-1)!}\mu^{a+1}(1-\mu)^{b-1} d\mu \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\mathrm{B}(a+2,b) \\
&= \frac{(a+b-1)!}{(a-1)!(b-1)!}\frac{(a+1)!(b-1)!}{(a+b+1)!} = \frac{a(a+1)}{(a+b+1)(a+b)}
\end{aligned}$$

$$\mathbb{E}(\mu^3) = \frac{a(a+1)}{(a+b+2)(a+b+1)(a+b)}$$

a and b, control the values of all moments of the distribution


b
10
5
0.2
st.dev.
0.1
0.0
5
a
10


10
b
5
1
skewness
0
−1
5
a
10

posterior pdf

$$p(\theta|n,N) = \mathrm{Beta}(\theta|n+1, N-n+1)$$

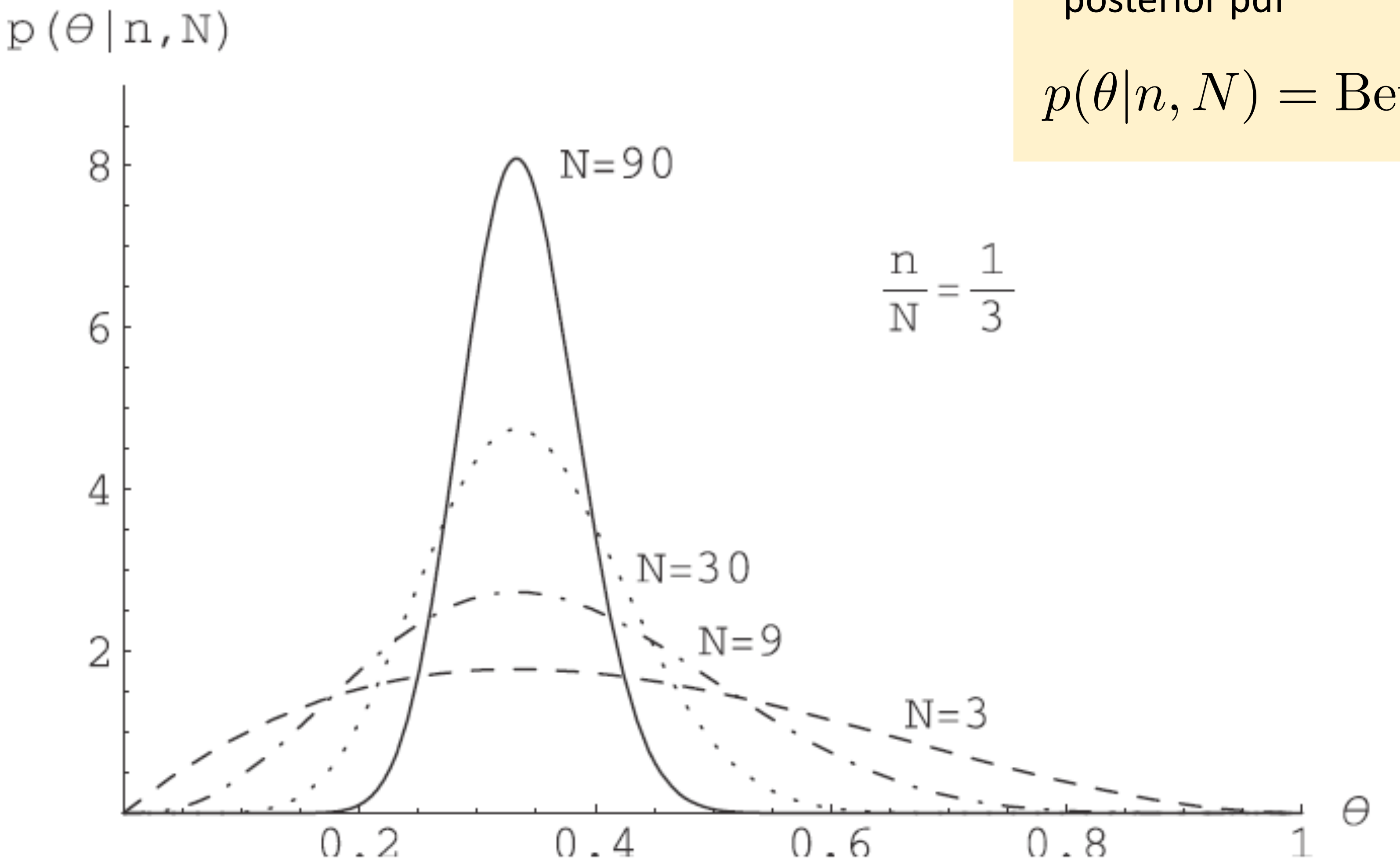


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

From the knowledge of the posterior pdf we find

$$p(\theta|n, N) = \mathrm{Beta}(\theta|n+1, N-n+1)$$

$$\begin{cases} n = a - 1 & \text{number of 1's} \\ N - n = b - 1 & \text{number of 0's} \end{cases}$$

meaning of the hyperparameters

$$\mathbb{E}(\theta) = \frac{n+1}{N+2}$$

biased, asymptotically unbiased, estimator

$$\mathrm{var}(\theta) = \frac{(n+1)(N-n+1)}{(N+3)(N+2)^2}$$

Maximum a posteriori (MAP) estimate – MAP ≠ mean value!

Consider the case with a uniform prior: from the posterior distribution

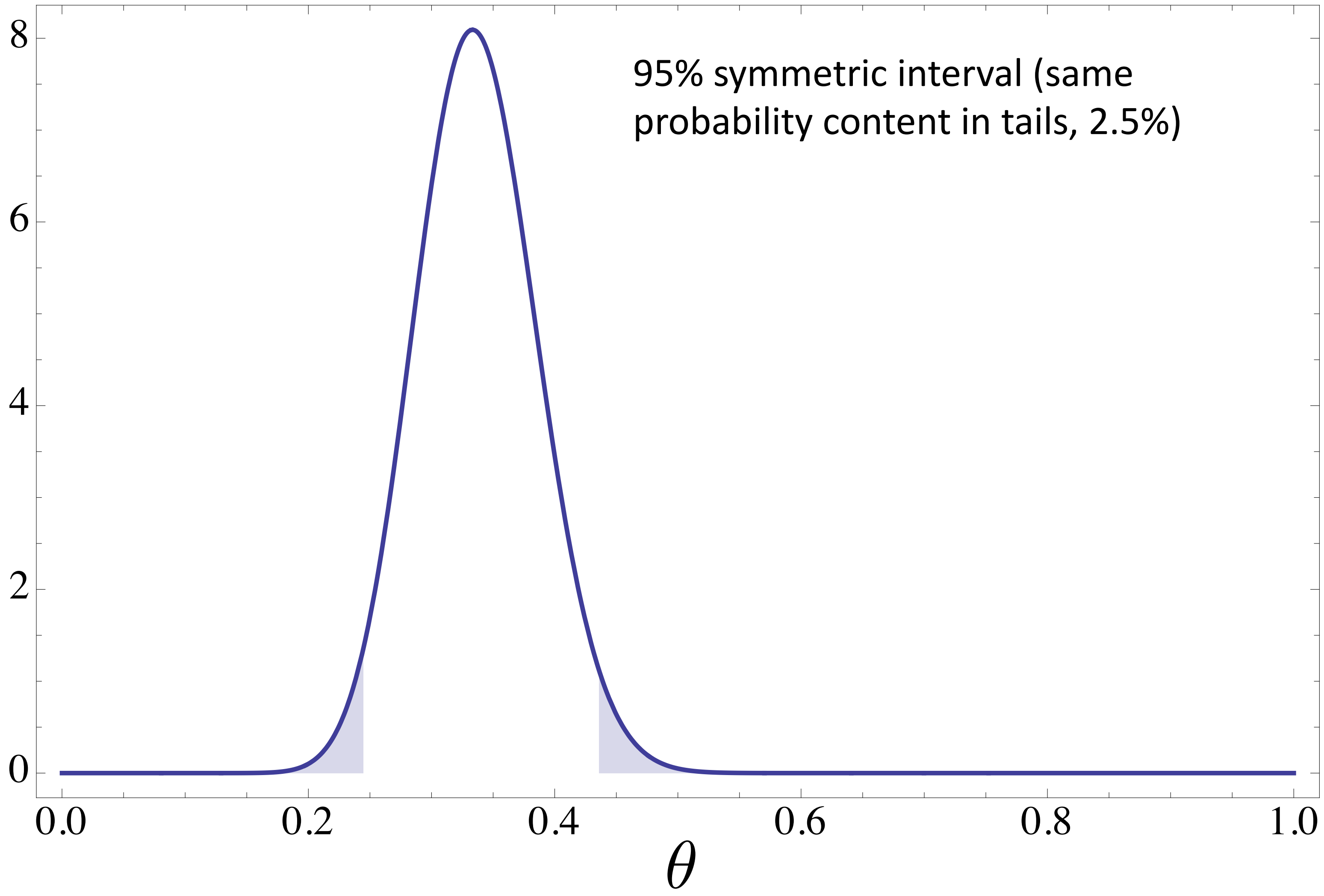
$$p(\theta|n,N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\theta^n$$

we easily find that the posterior pdf is maximized by the parameter value

$$\theta = n/N$$

which is the unbiased estimate of the parameter (unlike the mean value!)

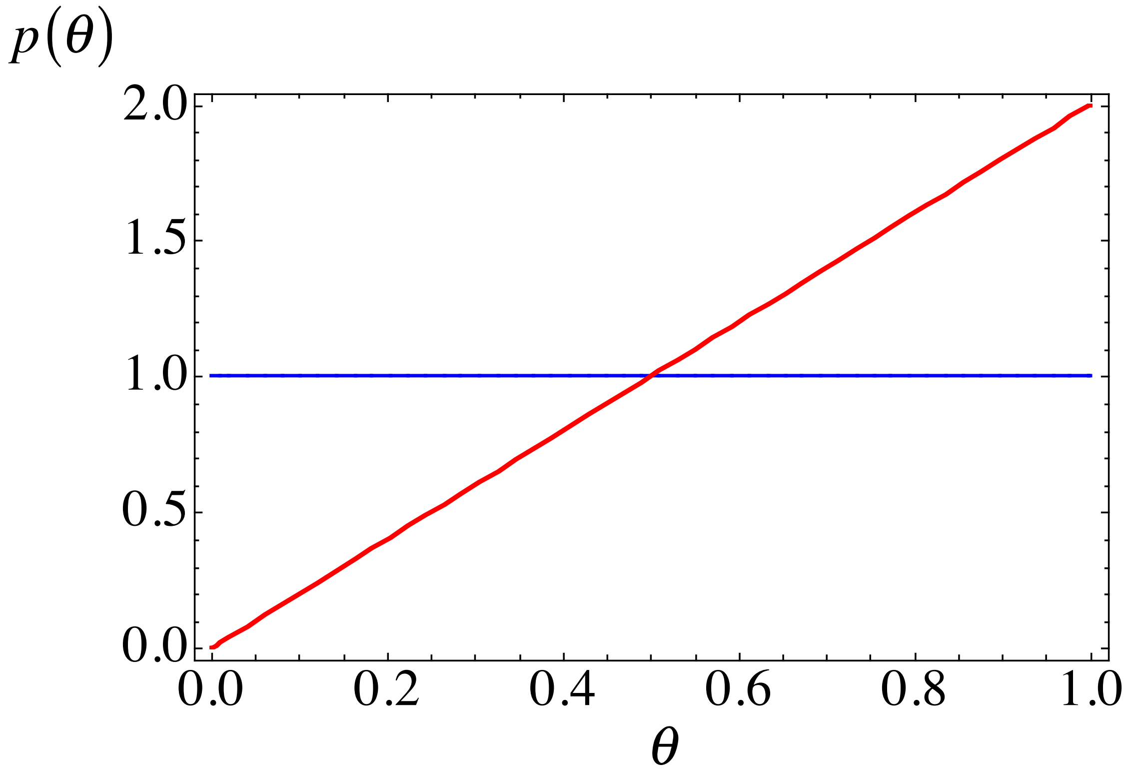
Credible intervals (case of initial uniform prior), the Bayesian analog of confidence intervals.



What happens if we try a different prior?

Let's try with a linear prior

$$p(\theta) = 2\theta$$



$$p(\theta|n,N) = \frac{P(n|\theta,N)}{\int_0^1 P(n|\theta',N)p(\theta')d\theta'}\, p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\; 2\theta' d\theta'}\, 2\theta = \frac{(1-\theta)^{N-n}\theta^{n+1}}{\int_0^1 (1-\theta)'^{N-n}\theta'^{n+1}\; d\theta'}$$

a beta distribution … again …

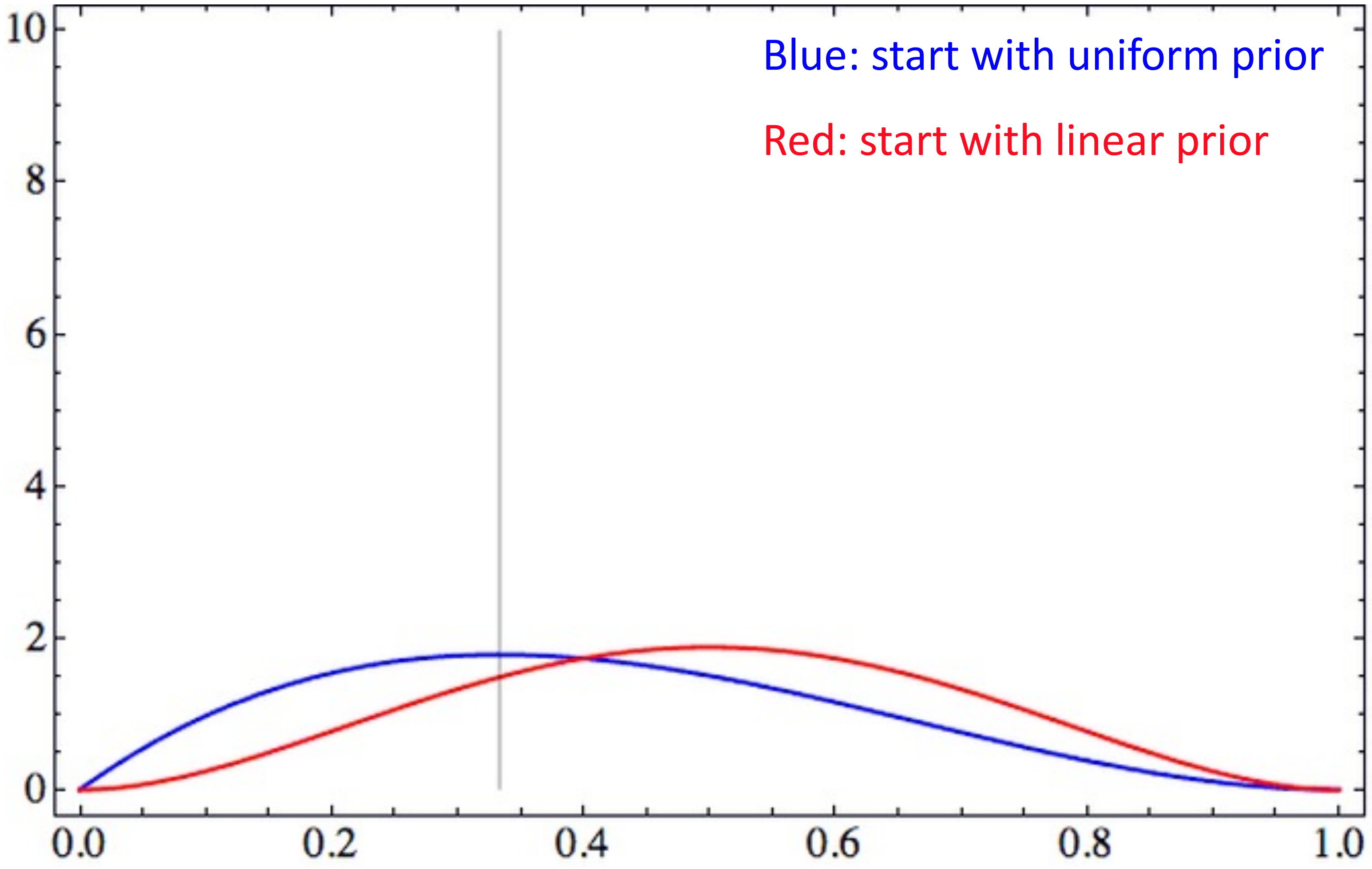
Properties of the new posterior

$$p(\theta|n, N) = \text{Beta}(\theta|n+2, N-n+1)$$

$$\begin{cases} n = a - 2 & \text{effective number of 1's} \\ N - n = b - 1 & \text{effective number of 0's} \end{cases}$$

$$\mathbb{E}(\theta) = \frac{n+2}{N+3}$$

$$\text{var}(\theta) = \frac{(n+2)(N-n+1)}{(N+4)(N+3)^2}$$



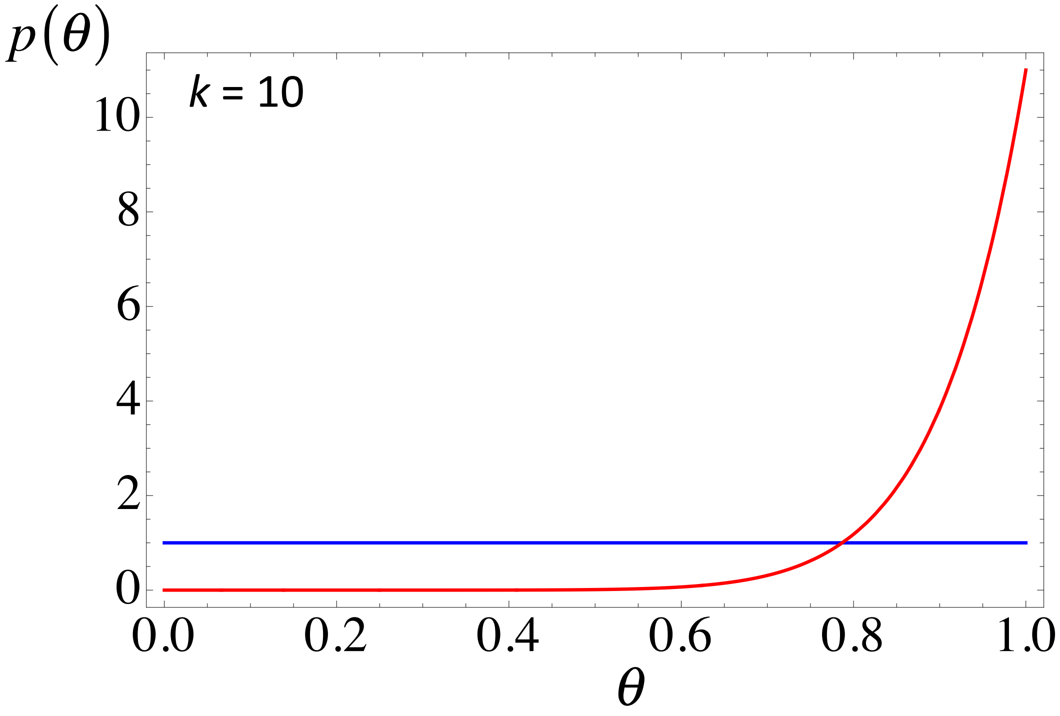
θ

Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

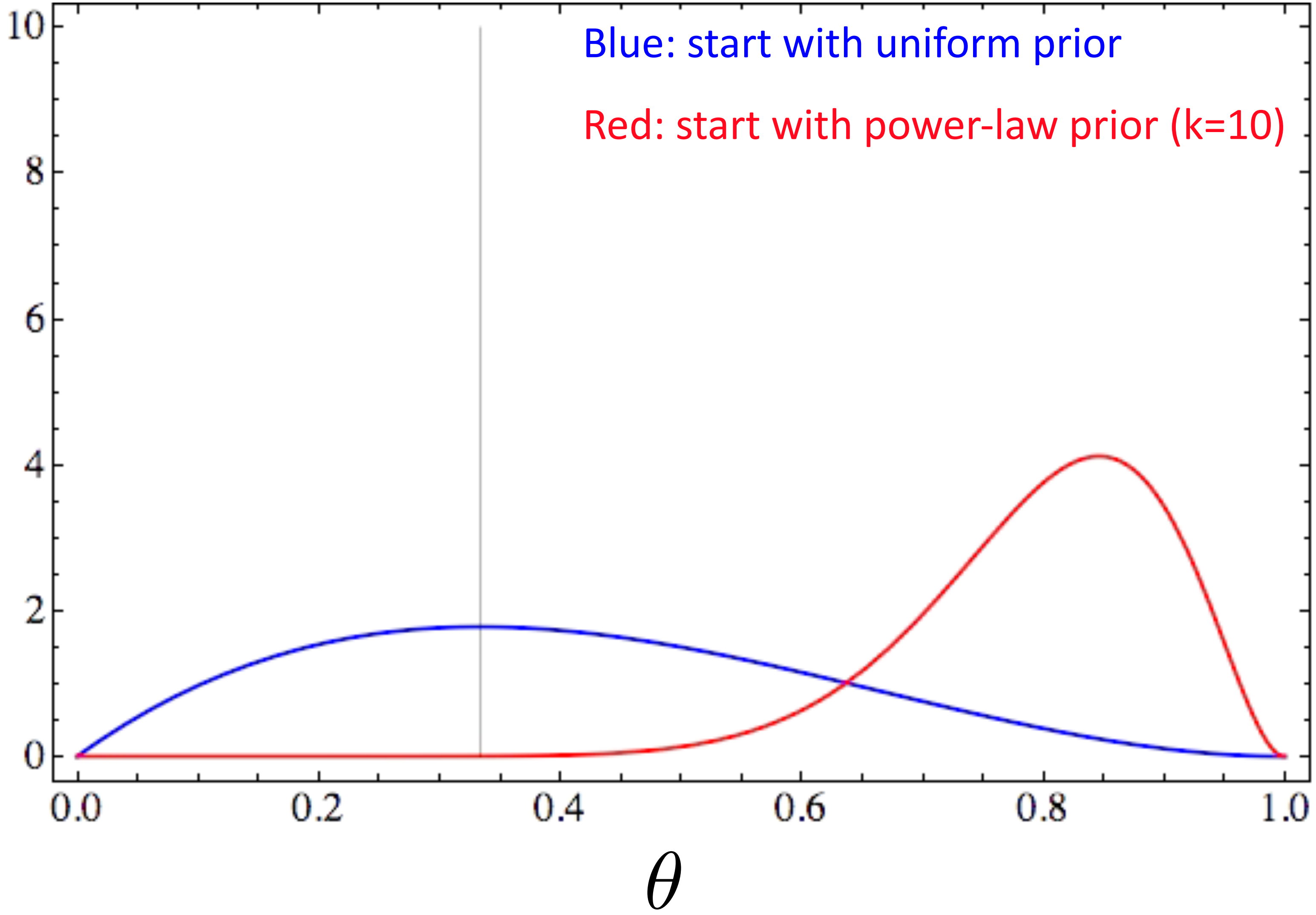
$$p(\theta) = (k+1)\theta^k; \quad k \gg 1$$



$$p(\theta|n,N) = \frac{P(n|\theta,N)}{\int_0^1 P(n|\theta',N)p(\theta')d\theta'}\, p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)'^{N-n}\theta'^n\;(k+1)\theta'^k d\theta'}\;(k+1)\theta^k = \frac{(1-\theta)^{N-n}\theta^{n+k}}{\int_0^1 (1-\theta)'^{N-n}\theta'^{n+k}\; d\theta'}$$

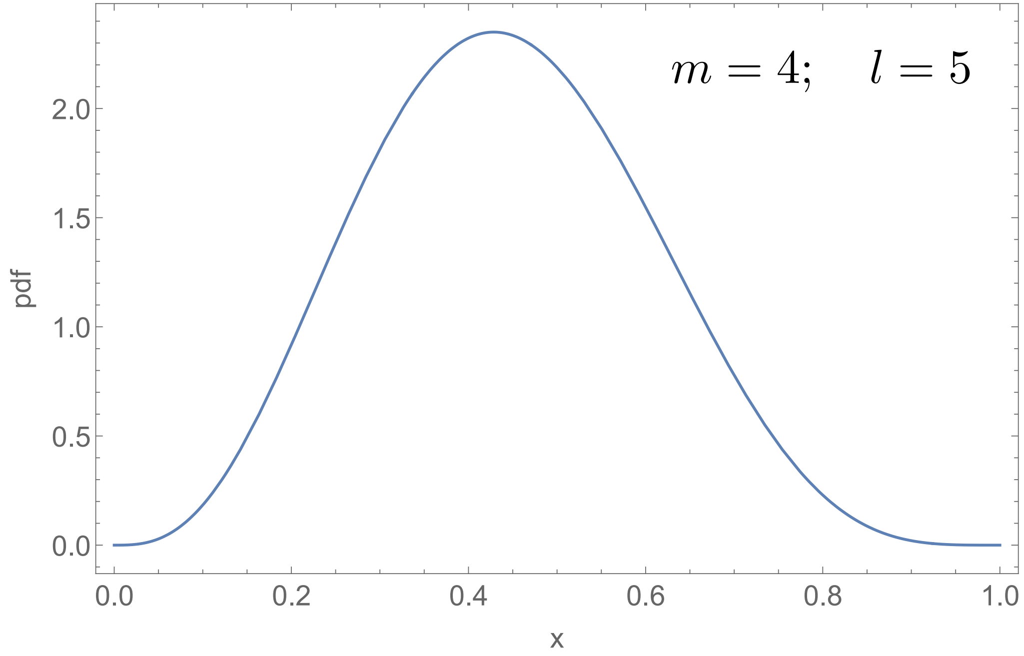
a beta distribution … yet again …



In this case, initial bias due to the prior is very large.

More generally, we can take a Beta prior

$$p(\theta) = \mathrm{Beta}(\theta|m,l) \propto \theta^{m-1}(1-\theta)^{l-1}$$



$$p(\theta|n,N) = \mathrm{Beta}(\theta|n+m, N-n+l)$$

$$\begin{cases} n+m & \text{effective number of 1's} \\ N-n+l & \text{effective number of 0's} \end{cases}$$

and

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{n+m}{N+m+l}$$

Lessons learned:

1. The prior information is not neutral, a careful choice of the prior distribution is a necessity.

   Question: how do we choose a prior?

2. If we want to keep all possibilities alive, we must heed the Cromwell's rule: "Prior probabilities 0 and 1 should be avoided" (Lindley, 1991)

   The reference is to Oliver Cromwell's phrase:
   I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

3. Convergence as the dataset size grows seems to be granted, however it may be very slow with a bad choice of prior distribution

   Question: is convergence really granted???

Lessons learned - 2:

1. While keeping the same functional form, the Binomial distribution for data leads to a Beta distribution for the parameter

2. Thanks to the functional shape of the likelihood, a Beta prior leads to a Beta posterior. Whenever a posterior has the same functional form as the prior, we call them *conjugate*.

3. Thanks to conjugacy, in this case we see that we can feed whole chunks of data to an existing posterior and still find that the sequential learning property of Bayes theorem is satisfied.

4. As the number of observations grows, the posterior is increasingly peaked → connection with frequentist statistics